ATLANTIS PRESS

Advances in Intelligent Systems Research, volume 153

International Conference on Applied Mathematics, Modeling and Simulation (AMMS 2017)

Real-time Haze Monitoring Based on Social Sensors

Dandan Zhu[1], Chenchen Wang[1,*], Dong Chen[2], Zhihui Ye[2] and Yuanfeng Lian[1]

[1]College of Computer Science, China University of Petroleum – Beijing, China

[2]College of Petroleum Engineering, China University of Petroleum – Beijing, China

[*]Corresponding author

***Abstract*—Traditional environmental monitoring systems require distributed sensors to collect data, which is restricted by equipment deployment. In contrast, social sensor, oriented by Social Network users, offers real-time, widely-distributed and low-cost sensory information gathered and processed through vision, hearing, touch et al. As described in this paper, we present a real-time monitoring system based on Weibo streams to deal with the prediction problem of Haze. The system is composed of two core modules: one is a classifier-based sensor signal indicator to select timely haze data, and the other is a temporal prediction model to forecast haze. The experimental results on the Weibo stream of October, 2016, illustrate high consistency with the corresponding AQI record, and imply that the proposed system is sensitive to the air quality change trend even in user inactive period.**

Keywords-event detection; hazing monitoring; classifier; temporal model; social sensor

I. INTRODUCTION

As environment issue is ever increasingly serious, environment monitor systems have to capture all kinds of paroxysmal events at any time, and thus, they are required of better immediacy and accuracy. Traditional environment monitor system is typically integrated by sensor network, embedded computation, modern network, wireless communications and distributed intelligent information disposal techniques.

Social Network Service (SNS) users around almost everywhere can perceive ambient information through their sensory organs such as eyes and ears, and then use their brains to screen out useful information. All these valuable data can be transferred in time to the platform of SNS and explosively accumulated. Regarding SNS users as world-wide distributed sensors for various information, their posts can be utilized as the sensor signals by proper selection. Compared with physical sensor, using social sensors for environment monitoring have unique advantages:

1. Large volume of data streams guarantee reliable sensor signals;
2. SNS users are movable and widely distributed, constructing a giant sensor network with ultra-wide coverage;
3. Social sensor network is heavily resilient, with many "sleeping" social sensors may be evoked at any time;
4. Social sensor network can be constructed without design and deployment;
5. Expert analysis on professional data, such as cloud atlas, is not required.

Above all, social sensors are suitable for constructing large-scale remote monitoring network system. By keeping a high level of activeness, Weibo streams carry large and diverse amount of information about latest news and events. And thus, it can be considered a valuable resource for detecting events which is defined as real-world occurrences that unfold over space and time [1, 2, 3], for instance, earthquake and typhoon.

As a dynamic source of information, Social Network streams, especially Twitter, have been attracting a lot of attention. Twitter has been proved to be useful for flu detection. Jiwei Li et al. [4] presented a Markov spatio-temporal model for monitoring Tweets. They built up a real-time flu reporting system to identify flu outbreaks and assist in emergency measure. Paul et al. [5] established the Ailment Topic Aspect Model (ATAM) which can isolate various ailments within a tweet corpus.

Therefore, the streams of the SNS contains rich event information, which enables a large amount of low-cost social sensors for event monitor. In this research, we chose Weibo streams to serve the air condition monitor systems. We adopted Weibo streams to establish a ubiquitous haze data acquisition network and design a real-time air quality monitoring system for haze prediction.

The contributions of this paper is summarized as follows:

1. Build a social sensor signal indication mechanism for time-limited haze data selection;
2. Design a user activation aware prediction model which alleviates the data sparse problem in certain time sections;
3. Propose a feasible construction scheme of ubiquitous monitoring network with early warning mechanism.

This article presents the whole monitor system building, as well as the real-world data test results. Section II gives the whole framework of the proposed system; section III and IV are the corresponding details of the main modules construction; experiments on real-world data have been conducted, and the results and analysis are shown in section V. Finally, we conclude our work in section VI, as well as the future work.

II. PROBLEM STATEMENT AND SOLUTION

We used the word "haze" as the query word to retrieve Weibo posts in Beijing from October 1st to 31st, 2016, and found that the daily quantity of the retrieved posts is roughly in line with the Air Quality Index records. The high correlation between

the number of Weibo posts and the AQI records implies that the Weibo streams carries information which can used for haze detection in the real world.

Consequently, we designed a real-time haze monitor system by analyzing the Weibo streams, which is capable of calculating the probability of haze occurrence and making early warning of haze. The whole system consists of 4 modules:

- Haze Weibo post filter: retrieve haze-related textual posts from Weibo streams by using "haze" as the query word;
- Data preprocessing: preprocess the Weibo returned from step one, including word segmentation, removal of stop words and punctuation, etc.;
- Sensor signal identification: classify data into signal or non-signal;
- Haze prediction: generate haze occurrence probability by using Weibo streams processed by the above steps.

The flow diagram of the proposed system is shown in figure Ⅰ.

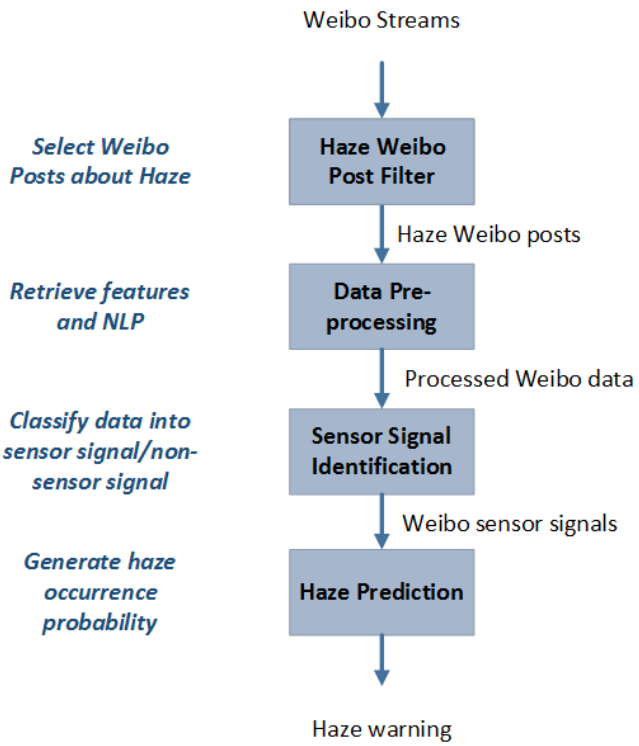


FIGURE I. HAZE MONITOR SYSTEM FRAMEWORK

III. SENSOR SIGNAL IDENTIFICATION

Social sensors are always very noisy compared to ordinal physical sensors. After retrieving the Weibo posts by the query word "haze", we preprocessed the raw data by a series of natural language processing(NLP) methods, including word segmentation, stop words removal, etc. However, not all the haze-related Weibo posts indicate the occurrence of a haze, and the system need to have the ability to identify the real sensor signal and filter non-sensor signal. Accordingly, a classifier-based sensor signal indicator has been created by a manually labeled haze corpus.

A. Labeling

In order to train the "signal/non-signal" classifier, we need to build a labeled haze corpus in which each document has a label to tell if it can be adopted as a sensor signal of the haze occurrence. Below are two examples of the "signal/non-signal" Weibo posts:

The haze came again! I could only stay at home. *(1)*

It didn't have a haze at this time last year. *(2)*

Apparently, we can draw a conclusion from sentence (1) that "haze is happening at sometimes, somewhere". However, sentence (2) is just describe an event out of date. So the former could be regarded as "sensor signal" and the latter belongs to "non-sensor signal".

After the natural language processing on the raw Weibo posts, we manually labeled each post with "1/0" according to "signal/non-signal".

B. Feature Selection

Feature selection can detect some irrelevant and redundant features and thus we used it to improve the efficiency and accuracy of the classifier. Multiple classifiers and text features are selected as control groups to get a better performance. Three groups of text-based features are employed in the experiments:

Feature A: the number of words in a Weibo content, and the position of the query word within a Weibo.

Feature *B*: the words before and after the query word

Feature *C*: all words in a Weibo

C. Classifier Building

SVM with a linear kernel, a Naïve Bayes classifier and a Random Forest classifier as effective classify models has been widely used in many natural language processing applications such as text categorization, information retrieval, etc. Consequently, we used the labeled corpus to train these three classifiers respectively so as to identify the sensor signals.

IV. HAZE PREDICTION MODEL

A. User Activation Degree

The users of Weibo show different activation degrees at different periods of time. The daily distribution of the obtained Weibo posts was illustrated in figure .Ⅱ By observing the quantity of Weibo separated by hours, we found especially from 3:00 to 4:00, which is the deep sleeping time for person, the quantity is nearly zero.

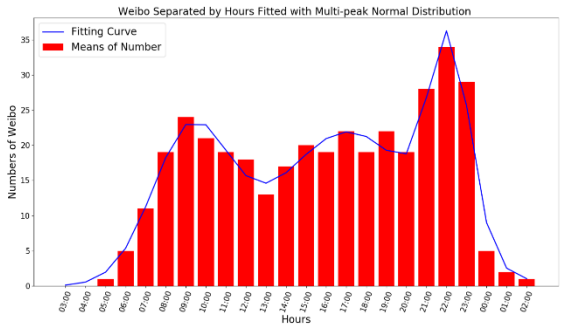


FIGURE II. NUMBERS OF WEIBO AND ENVIRONMENTAL QUANLITY REPORT

To describe the activity of Weibo, we quoted "ω" as the user activation degree like following:

$$\omega = Af(t) \tag{1}$$

"ω" means the user activation index and will get improved as the activity of Weibo increasing where $f(t)$ is a multi-peak normal distribution. "A" is a uniform coefficient to make the user activation index in an adequate range.

B. Temporal Prediction Model

By observing the quantity of Weibo about haze, which fit very well to the exponential distribution. We assumed that every social sensor (a Weibo terminal) is independent and identically distributed. We can infer that if a user detects event at time 0, assume that the probability of his posting a Weibo from t to Δt is fixed as λ. Supposing we got N_0 sensor signals at time 0, the sum of sensor signals from time 0 to time t we will get is $\frac{N_0(1-e^{-\lambda(t+1)})}{(1-e^{-\lambda})}$.

By introducing the user activation degree, we designate a score function to describe of the amendatory sensor signal amount at time t:

$$(t) = \frac{1}{\omega}\left(\frac{N_0(1-e^{-\lambda(t+1)})}{(1-e^{-\lambda})} + C\right) \tag{2}$$

where the user activation degree ω is used as a balance factor in order to eliminate the effects of the user activation, and the constant C is a smoothing parameter.

Since the sensor signal indicator may misjudge a Weibo post as a sensor signal, we model the probability of the haze occurrence at time t as follows:

$$P_{occur}(t) = 1 - P_f^{\frac{1}{\omega}\left(\frac{N_0(1-e^{-\lambda(t+1)})}{(1-e^{-\lambda})}+C\right)} \tag{3}$$

where P_f is the misjudge-ratio of a sensor.

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A. Corpus Generation

We gathered Weibo posted in October and whose location like "Beijing" by the spider programs. The size of original data set we have collected is 11745. After NLP and labeling, we obtained the labeled haze corpus which size is 5000.

B. Sensor Signal Identification Results

The labeling corpus was divided into the training set (80%) and the test set (20%) in order to train and test a classifier. In the experiments, we compared MultinomialNB, LinearSVC and RandomForest Classifier to get a better performance.

Testing different classifiers respectively based on feature C, the results are shown at table I. The results inferred that the MultinomialNB classifier performs the best. The MultinomialNB got the highest value which were more than 80 percent at the four standards containing accuracy, recall, F1 score and precision. But the others were tested differ hugely in results which didn't exceed 76%.

Thus we chose MultinomialNB as the classifier in sensor signal identification module and then we tested different text features combinations based on it. The scores are illustrated at Table Ⅲ.The results showed that there is little difference between features A+B+C and only C on the four standards, so the contributions of feature A and B were poor and the performance of feature C was best. From the test results of feature A and B separately, the scores were too low to take them consider as the features.

So considering computational costs in the experiments, it is suitable to select MultinomialNB classifier based on feature C as the central section of signal identification module.

TABLE I. CLASSIFIER PERFORMANCES ON FEATURE C

Classifiers	Accuracy	Recall	F1 score	Precision
MultinomialNB	82.00%	80.50%	81.73%	82.99%
LinearSVC	70.00%	69.00%	69.70%	70.41%
RandomForest Classifier	75.75%	75.50%	75.69%	75.88%

TABLE II. MULTINOMIALNB CLASSIFIER PERFORMANCE ON DIFFERENT FEATURES

Text features	Accuracy	Recall	F1 score	Precision
A	59.75%	60.54%	61.42%	62.30%
B	67.00%	68.90%	69.60%	70.30%
C	82.00%	80.50%	81.73%	82.99%
A+B+C	79.50%	80.07%	80.32%	80.57%

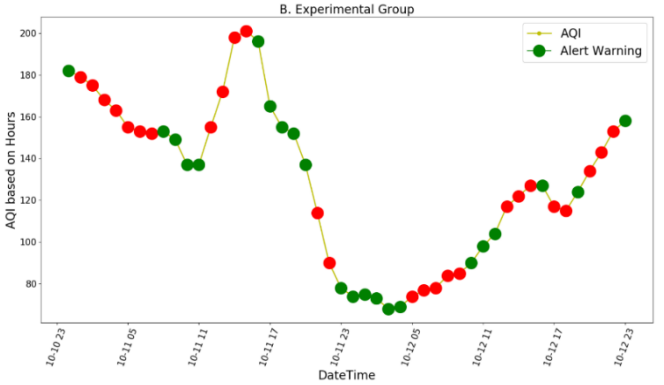
C. Haze Prediction Results

In the experiments, we chose methods proposed by Sakaki, et al.[1] as the control group which is the baseline of the proposed prediction model.

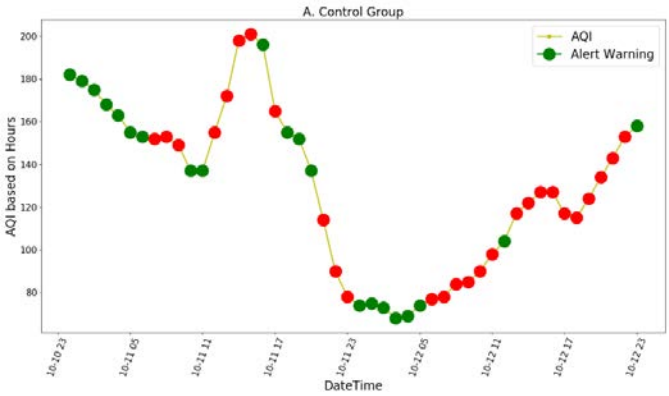
We extracted the Weibo streams from October 11th to October 12th to make a real-world test of the proposed system. During the experiment, we built a haze alarm by setting a threshold for $P_{occur}(t)$, that is, $P_{threshold}$ = 95%. Under the alarm mechanism, the system will alert when the predicted probability of haze occurrence, $P_{occur}(t)$, reaches $P_{threshold}$. Based on the scores of testing MultinomialNB classifier, we set P_f=0.2, A=0.5 and C=0.1.

Figure III represents the alert situation based on the control group and the proposed system. The horizontal axis in the graph

is the testing time, and the vertical axis is the real AQI value of the corresponding time. The red and green dots on the curve represent the alert results. The red dots means the system is starting the haze alarm while the green dots means there is no alarm.



A. Control Group



B. Experimental Group

FIGURE III. HAZE PREDICTION PERFORMANCES

Comparing the alarm status with the real AQI data, we can find out the control group and proposed system could accurately judge the quality of air condition in the high user activation periods (from 7:00 a.m. to 24:00 p.m. in October 11th and 12th) and alerted when the air quality was bad. But in the low user activation periods (from 0:00 to 6:00 a.m. in October 11th and 12th) because of the sharp fall in Weibo streams, the control group didn't receive enough sensor signals to start the alarm. So in that time, the control group was getting into a dormant period and unable to detect the air quality. While the proposed system can enlarge the amount of sensor signals and was more sensitive in inactive phase because of the introduction of score function. As the graph B in Figure III shows, the AQI value was high and the proposed system accurately alarmed from 0:00 to 6:00 a.m. in October 11th, but the control group in graph A had no alarm.

V. CONCLUSIONS AND FUTURE WORK

In this paper, we presented a social sensor based system to detect haze occurrence in real time. The experimental results indicated that the proposed system can make a valid judgement on haze occurrence and be sensitive to the activation of users.

In future, we will focus on the following points: to improve the accuracy of the sensor signal identification; to make the model simulating the process of haze happening, developing, peaking and descending accurately; to optimize the user activation function and temporal model, and test the model based on a larger training set.

ACKNOWLEDGEMENTS

This study was financially supported by Science Foundation of China University of Petroleum, Beijing (No.2462015YJRC008).

REFERENCES

[1] Sakaki, et al. "Earthquake shakes Twitter users: real-time event detection by social sensors." (2010).

[2] Allan, James, et al. "Topic Detection and Tracking Pilot Study Final Report." Darpa Broadcast News Transcription and Understanding Workshop 1998:194--218.

[3] Yang, Yiming, T. Pierce, and J. Carbonell. A study of retrospective and on-line event detection. Proceedings of the 21st Annual international ACM SIGIR Conference on Research and Development in information Retrieval. SIGIR '98. 1998.

[4] Li, Jiwei, and C. Cardie. "Early Stage Influenza Detection from Twitter."

[5] Computer Science (2013).

[6] M. Paul and M. Dredze. A model for mining public health topics from twitter. HEALTH, 11:16-6, 2012.

[7] Atefeh, Farzindar, and W. Khreich. "A Survey of Techniques for Event Detection in Twitter." Computational Intelligence 31.1(2015):132-164.

[8] Salton, Gerard. "Automatic text processing: the transformation, analysis, and retrieval of information by computer." (1989).

[9] Kumaran, et al. "Text classification and named entities for new event detection." 20.17(2004):297-304.

[10] SNOWSILL, T., F. NICART, M. STEFANI, T. DE BIE, and N. CRISTIANINI. 2010. Finding surprising patterns in textual data streams. In Cognitive Information Processing (CIP), 2010 2nd International Workshop, Tuscany, Italy, pp. 405–410.

[11] SNOWSILL, T., F. NICART, M. STEFANI, T. DE BIE, and N. CRISTIANINI. 2010. Finding surprising patterns in textual data streams. In Cognitive Information Processing (CIP), 2010 2nd International Workshop, Tuscany, Italy, pp. 405–410.

[12] FUNG, G. P. C., J. XU YU, P. S. YU, and H. LU. 2005. Parameter free bursty events detection in text streams. In Proceedings of the 31st International Conference on Very Large Data Bases, VLDB '05, VLDB Endowment, pp. 181–192.

[13] HE, Q., K. CHANG, and E.-P. LIM. 2007. Analyzing feature trajectories for event detection. In Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '07, ACM, New York, NY, pp. 207–214.

[14] HE, Q., K. CHANG, E.-P. LIM, and J. ZHANG. 2007. Bursty feature representation for clustering text streams. In SIAM International Conference on Data Mining.

[15] SNOWSILL, T., F. NICART, M. STEFANI, T. DE BIE, and N. CRISTIANINI. 2010. Finding surprising patterns in textual data streams. In Cognitive Information Processing (CIP), 2010 2nd International Workshop, Tuscany, Italy, pp. 405–410.